基于混沌 PSO 的高维多视图数据 IWKM 聚类算法 *

陈高祥 [1,2]，陈都鑫 [1]

(1. 东南大学 计算机科学与工程学院，南京 211189; 2. 江苏联合职业技术学院 苏州学院，江苏 苏州 215000)

摘　要：针对传统聚类算法无法处理大数据中多视图高维数据问题，提出了一种基于混沌粒子群优化算法的智能加权 K 均值聚类算法。首先，在聚类模型中引入聚类之间的耦合程度以扩大聚类的相似性。其次，为了消除初始聚类中心的敏感性，利用混沌粒子群优化算法通过全局搜索得到最优初始聚类中心、视图权重和特征权重。然后，引入一种精确摄动策略提高混沌粒子群优化算法的寻优性能。最后通过在 apache spark 和 single node 两个平台上的实验验证了提出的方法在视图多、维数高的复杂数据集条件下具有较好的聚类性能。

关键词：大数据；K 均值聚类；高维多视图数据；粒子群优化算法

中图分类号：TP301.6　　doi: 10.19734/j.issn.1001-3695.2020.02.0045

IWKM clustering algorithm for high dimensional multi view data based on chaos PSO

Chen Gaoxiang[1,2], Chen Duxin[1]

(1. *School of Computer Science & Engineering*, *Southeast University*, *Nanjing Jiangsu* 211189, *China*; 2. *Suzhou College*, *Jiangsu United Vocational & Technical College*, *Suzhou Jiangsu* 215000, *China*)

Abstract: Aiming at the problem that traditional clustering algorithm can't deal with multi view and high dimension data in big data, this paper proposed an intelligent weighted K-means clustering algorithm based on chaos particle swarm optimization algorithm. Firstly, it introduced the coupling degree between clusters to expand the similarity of clusters. Secondly, in order to eliminate the sensitivity of the initial clustering center, it used chaos particle swarm optimization algorithm to obtain the optimal initial clustering center, view weight and feature weight through global search. Then, it introduced an accurate perturbation strategy to improve the performance of chaos particle swarm optimization. Finally, experiments on Apache spark and single node show that the proposed method has better clustering performance under the condition of complex data sets with multiple views and high dimensions.

Key words: big data; K-means clustering; high-dimensional multi view data; particle swarm optimization algorithm

0　引言

当前，人工智能、移动互联网、社交网络和物联网生成大量数据，并推动大数据应用的快速发展[1,2]。在各种大数据应用中，都有许多高维多视图数据，高维多视图数据通常以各种来源获得的多个特征空间和不同结构进行描述[3]。传统聚类方法将所有视图作为一个统一的变量集，对此类具有数量、种类、速度、准确性和价值等多视图的数据聚类效果较差[4]。在大数据环境下，如何实现高维多视图数据的聚类以适应各种复杂的、大规模的应用是最具挑战性的问题之一[5]。

由于大数据的广泛应用，多视图数据的聚类吸引了许多研究人员的关注。文献[6]将多视图任务的非监督特征选择保留在集群结构中，然后提出一种交替算法来实现该结构。对于多视图聚类问题，文献[7]提出了一种新颖的多视图关联传播算法，该算法特别适合于对两个以上的视图进行聚类。在文献[8]中，提出了局部自适应聚类(local adaptive clustering, LAC)算法，该算法为每个聚类的每个特征分配权重，通过使用迭代算法最小化其目标函数。文献[9]等提出了一种多视图数据的自动两级变量加权 K 均值(two-variables weighted K-means, TWKM)聚类算法，该算法可以同时计算视图和单个变量的权重，但是很容易导致在单个特征和单个视图上具有较大权重的聚类，因此权重的分布不平衡。然而，上述算法主要关注于具有视图方式关系的问题，而忽略了数据集高维特征的重要性，使得聚类结果与实际应用存在较大差异，此外，上述方法由于聚类性能的限制对大数据应用中更复杂的高维数据的聚类效果较差。

为了解决实际大数据应用中的高维多视图数据聚类问题，并且进一步提升该聚类算法的聚类性能，提出了一种基于混沌粒子群优化算法(chaos particle swarm optimization, CPSO)的智能加权 K 均值(intelligent weighted K-means, IWKM)聚类算法。

1　相关理论

1.1　加权 K 均值聚类算法

集群是数据对象的集合，这些数据对象在同一集群中彼此相似，但与其他集群中的对象不同[10]。给定数据对象集 $X=[x_{i,j}]_{N*D}$，N 是数据对象的数量，D 是数据对象的维度。也就是说，数据对象具有 D 个特征。聚类问题试图找到 X 的 k 分区。簇的中心是 $Z=[z_{k,j}]_{C*D}$。$\boldsymbol{U}=[u_{i,k}]_{N*C}$，模糊除法矩阵，描述对象是某些集群的隶属度。

作为具有敏感初始聚类中心的聚类算法，K 均值被广泛用于实际应用中，例如图像分割和数据挖掘[11,12]。K 均值的目标是找到一个分区，以最小化带簇的平方和。在聚类过程中，用以下式子解决样本划分任务：

$$\begin{aligned}&\mathrm{F}(\boldsymbol{U},\boldsymbol{Z})=\sum_{k=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{D}u_{i,k}\left(x_{i,j}-z_{k,j}\right)^{2}\\&s.t.\sum_{k=1}^{C}u_{i,k}=1, 1\le i\le N, u_{i,k}\in\{0,1\}\end{aligned}\tag{1}$$

其中：$\boldsymbol{U}$ 被定义为分区矩阵，$u_{i,k}$ 是一个二进制变量。

收稿日期：2020-02-23；**修回日期：**2020-04-29　**基金项目：**国家自然科学基金资助项目(61903079)

作者简介：陈高祥(1978.1)，男，江苏南京人，讲师，硕士，主要研究方向为计算机应用(lzz701018@sina.com)；陈都鑫(1983.8-)，男，湖北武汉人，副教授，博士，主要研究方向为大数据处理.

$Z=\{Z_1,Z_2,\ldots,Z_k\}$ 是一组向量，表示 k 个簇的质心。$(x_{i,j}-z_{k,j})^2$ 是第 i 个对象与第 j 个变量上的第 k 个簇的中心之间的距离度量。

在经典的 K 均值聚类算法中，所有特征均具有相同的权重，在诸如消费者细分之类的聚类问题中，所有特征均得到同等对待。实际上，在许多实际应用中，数据集中不同特征对聚类的影响是不同的，因此有必要为不同特征分配不同的权重。K 均值类型聚类中的自动变量加权是加权 K 均值聚类算法，目标函数为

$$\mathrm{F}(\boldsymbol{U},\boldsymbol{Z},WF)=\sum_{k=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{D}u_{i,k}wf_j^{\beta}\left(x_{i,j}-z_{k,j}\right)^2$$
$$s.t.\sum_{k=1}^{c}u_{i,k}=1,,u_{i,k}\in\{0,1\}\quad\sum_{j=1}^{D}wf_j=1,0\le wf_j\le 1 \tag{2}$$

其中，$\boldsymbol{U}$ 被定义为 $n\times k$ 分区矩阵。 WF 是特征的权重。

1.2 软子空间聚类算法

软子空间聚类算法根据维度在发现相应聚类中的作用来确定维度的子集。维度的贡献是通过在聚类过程中分配给维度的权重来衡量的。在文献[13]中提出了一种软子空间聚类算法，目标函数的建模为

$$\mathrm{F}(\boldsymbol{U},\boldsymbol{Z},\boldsymbol{WCF})=\sum_{k=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{D}u_{i,k}wcf_{k,j}^{\beta}\left(x_{i,j}-z_{k,j}\right)^2$$
$$s.t.\sum_{k=1}^{C}u_{i,k}=1,1\le i\le N,u_{i,k}\in\{0,1\},$$
$$\sum_{j=1}^{D}wcf_{k,j}=1,0\le wcf_{k,j}\le 1 \tag{3}$$

其中 $\boldsymbol{WCF}$ 是每个集群中每个属性的权重。

2 IWKM 算法

2.1 高维多视图数据的聚类模式

用于将 x 划分为具有视图和特征权重的集群的聚类建模为以下目标函数的最小化。

$$\min\ \ \mathrm{Fitness}(\boldsymbol{U},\boldsymbol{Z},\boldsymbol{WV},\boldsymbol{WF})=\frac{\sum_{k=1}^{C}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j\in\mathrm{View}_t}u_{i,k}wv_t wf_j\left(x_{i,j}-z_{k,j}\right)^2}{\sum_{k=1}^{C}\sum_{t=1}^{T}\sum_{j\in\mathrm{View}_t}wv_t wf_j\left(z_{k,j}-o_j\right)^2}$$
$$s.t.\begin{cases}\sum_{k=1}^{C}u_{i,k}=1,1\le i\le N,u_{i,k}\in[0,1]\\ \sum_{t=1}^{T}wv_t=1,0\le wv_{Pt}\le 1\\ \sum_{j\in View_t}wf_j=1,0\le wf_j\le 1,0\le t\le T\\ po_j=\sum_{k=1}^{C}z_{k,j}/C\end{cases} \tag{4}$$

其中 $\boldsymbol{U}=[u_{i,k}]_{N\times C}$ 是一个 $N\times C$ 分区矩阵，其元素 $u_{i,k}$ 为二进制，其中 $u_{i,k}=1$ 表示对象 i 已分配给集群 k 。 $\boldsymbol{Z}=[z_{k,j}]_{C\times D}$ 是一个 $N\times C$ 矩阵，其元素 $z_{k,j}$ 表示簇 k 的第 j 个特征。 $\boldsymbol{WV}=[wv_t]_T$ 是 T 视图的权重。 $\boldsymbol{WF}=[wf_j]_j\in\mathrm{View}_t$ 是视图 t 下的特征权重。 $wv_t wf_j(x_{i,j}-z_{k,j})^2$ 是第 i 个对象与第 k 个簇的中心之间的第 j 个特征的加权距离度量。 $wv_t wf_j(z_{k,j}-o_j)^2$ 是第 k 个聚类与平均聚类中心之间的第 j 个特征的加权距离度量， o_j 是 C 个聚类的平均聚类中心。该值描述集群之间的耦合程度，越大表示相异性越大。

2.2 CPSO 和粒子编码

在 IWKM 中，提出了 CPSO 以帮助算法获得更好的初始聚类中心、视图权重和特征权重。每个粒子 i 代表 D 维解空间中的候选解，它具有两个向量：位置向量 $\boldsymbol{X}_i=[x_i^1,x_i^2,\ldots,x_i^D]$ 和速度向量 $\boldsymbol{V}_i=[v_i^1,v_i^2,\ldots,v_i^D]$ 。在演化过程中，通过以下等式更新迭代 $t+1$ 上维度为 d 的粒子 i 的速度矢量和位置矢量：

$$v_i^d(t+1)=\omega v_i^d(t)+c_1r_1(pBest_i^d(t)-p_i^d(t))$$
$$+c_2r_2(gBest^d(t)-p_i^d(t)) \tag{5}$$

$$x_i^d(t+1)=x_i^d(t)+v_i^d(t+1) \tag{6}$$

其中 $d=1,2$ ， D 表示搜索空间的每个维度， ω 是惯性权重， c_1 和 c_2 分别是认知学习系数和社会学习系数， r_1 和 r_2 是在 [0,1] 范围内的两个均匀随机数， $pBest_i^d(t)$ 是在粒子 i 的第 t 次迭代之前找到的具有最佳适应度的维度 d 上的位置， $gBest^d(t)$ 是整个粒子群在维度 d 上找到的最佳位置。惯性权重 ω 通常更新为

$$\omega=\omega_{\max}-(\omega_{\max}-\omega_{\min})\times g/g_{\max} \tag{7}$$

其中 $\omega_{\max}$ 和 $\omega_{\min}$ 是初始权重和最终权重，分别设置为 0.9 和 0.4。 g 是当前进化世代数， $g_{\max}$ 是最大世代数，并设置为 150。 c_1 和 c_2 分别设置为 1.8。维度 d 上每个粒子的速度限制在 $[-v\max^d,v\max^d]$ ， $v\max^d\in R^+$ 。因此，如果速度 $v_i^d(t)$ 超过 $v\max^d$ ，则将其重新分配给 $v\max^d$ 。否则，如果速度 $v_i^d(t)$ 小于 $-v\max^d$ ，则将其重新分配给 $-v\max^d$ 。如果 $v\max^d$ 太大，粒子可能会错过良好的解决方案。另一方面，如果 $v\max^d$ 太小，粒子可能会陷入局部最优状态。通常将最大速度 $v\max^d$ 设置为搜索范围的 20％。

粒子编码是使用粒子群搜索最佳解的前提[14]。在 IWKM 中，初始聚类中心、视图权重和特征权重被编码为粒子表示形式。每个粒子由 $F\times C+T+F$ 维实数向量编码。 F 是聚类问题中对象的特征数。群中的第 i 个粒子被编码为

$$\boldsymbol{X}_i=\begin{bmatrix}x_i^{1,1},x_i^{1,2},\ldots,x_i^{1,F},\ldots,x_i^{C,1},x_i^{C,2},\ldots,x_i^{C,F}\\ wv_i^1,\ldots,wv_i^T,wf_i^1,\ldots,wf_i^F\end{bmatrix} \tag{8}$$

$$x(t+1)=r\cdot x(t)\cdot(1-x(t)),r\in N,x(0)\in[0,1] \tag{9}$$

$$d(pBest,gBest)=\frac{1}{N_S}\sum_{i=1}^{N_S}\sqrt{\sum_{j=1}^{\dim}(pBest_{ij},gBest_j)^2}\le Q_d \tag{10}$$

$$sd(pBest_j)=\sqrt{\frac{1}{N_S}\sum_{i=1}^{N_S}\left(pBest_{ij}-\frac{1}{N_S}(pBest_{1j}+pBest_{2j}+\cdots+pBest_{N_sj})\right)^2}\le Q_pBest \tag{11}$$

$$sd(gBest_j)=\sqrt{\frac{1}{m-1}\sum_{t=1}^{m-1}(gBest_j(m)-gBest_j(t))^2}\le Q_gBest \tag{12}$$

2.3 精确扰动和 CPSO

为了避免局部最优和过早收敛，利用跳跃或突变在丰富群体智能中粒子的搜索行为方面具有很大的优势[15]。在 CPSO 中，混沌逻辑序列扰动被用于帮助粒子脱离局部最优并获得更好的搜索质量，具有确定性、遍历性和随机性，将其定义为等式(9)，其中 r 是控制参数， x 是变量， $r=4$ ， $t=0,1,2,\ldots$ 。

可以将 CPSO 的精确扰动概括为以下过程的相互作用。

a) 创建合适的扰动粒子：为了减少粒子搜索过程中总体稳定性的损害和 CPU 的计算负荷，通过简单的随机抽样方法从总共 N_S 个粒子中随机选择 N_S/K_spark 个粒子作为扰动对象。 K_spark 是 apache spark 中的工作程序节点数。

b) 精确扰动时间：扰动的时间是粒子群过早收敛的时间。 $pBest$ 和 $gBest$ 之间的平均距离用于判断粒子是否处于过早收敛状态，记,为等式(10)。其中， N_S 和 dim 是群的粒子数和粒子的维度， Q_d 代表过早收敛的阈值。如果 $d(pBest,gBest)\le Q_d$ ，则出现过早收敛和局部最优，然后应采用适合 N_S/K_spark 粒子的扰动。

c) 精确扰动维度：由于粒子具有一个以上的维，因此根据惯性的优先级，优先选择一些具有较高惯性的维来进行扰动。第 j 维中的 $pBest$ 和 $gBest$ 的惯性可以由均方差给出，分别记为等式(11)和(12)。其中 N_S 和 m 是群的粒子数和当前迭代数。如果 $sd(pBest_j)\le Q_pBest$ 或 $sd(gBest_j)\le Q_gBest$ ，则第 j 维的 $pBest$, $gBest$ 是惰性的，需要进行扰动。其中 Q_pBest 和 Q_gBest 分别是 $pBest$ 和 $gBest$ 的惰性阈值。

2.4　IWKM 流程

IWKM 算法的流程图如图 1 所示。

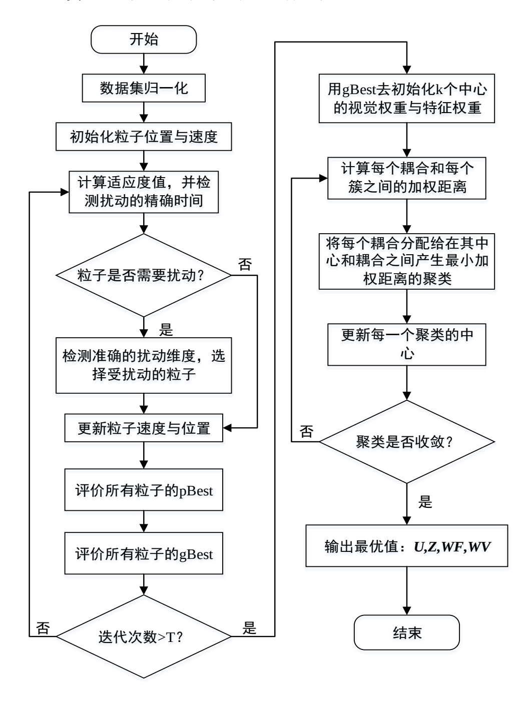


图 1　IWKM 算法流程图

Fig. 1　Flow chart of IWKM algorithm

3　实验评估

3.1　测试环境和 spark

在分布式和并行计算环境中，apache spark 是一个重要的开源集群计算框架，它为隐式数据并行和容错的整个集群编程提供了一个接口[16]。spark 的弹性分布式数据集(RDD)是分布式程序的工作集，可以提供受限形式的分布式共享内存。通常，由于重复的重新启动作业、大数据读取和改组，MapReduce 作为有效的海量并行数据处理框架，不适合迭代算法。因此，本文选择了 apache spark 作为大数据应用中 IWKM 的计算平台。在本文的实验中，对 IWKM 进行了测试，并在包括 apache spark 和 single node 在内的各种计算环境中进行了比较。Single node 配备了 Intel Core i5-4210M 2.6 Hz，3.8 G RAM 和 ubuntu 14.04LTS 操作系统。

Apache spark 由一个主节点，配置为 Intel Core i7-3820@3.6 GHz，64G DDRIII 和 1T 高效云磁盘，十个工作节点，相关配置为 Intel Xeon E5-2690@2.9 GHz，16G DDRIII 和 500G 高效云磁盘，应用版本为 apache spark 1.6.0。

3.2　测试数据集和评估指标

为了评估所提算法的性能，本文还通过 RI，JC 和 Folk 的评估指标，将 IWKM 与 LAC，亲和传播(affinity propagation,AP)[17]，归一化分割(normalized cut,Ncut)[11]，密度聚类(density clustering,DC)[18]和 TWKM 进行了比较。为了公平比较，PSO 和 CPSO 使用相同的人口规模 30 和相同的 150 个适应度值评估。IWKM 和其他五种算法已在 5 个高维多视图数据集中进行了测试，其中包括多特征数据集，互联网广告数据集，Spambase 数据集，图像分割集和心电图数据集。这些数据集及其应用的基本信息如表 1 所示。

表 1　高维多视图数据集的特征

Tab. 1　Characteristics of high dimensional multi view data set

序号	数据集	数据量	数据类别	数据特征	视觉数目	单个类的大小
1	多特征	2000	10	649	6	(200,200,...,200)
2	互联网广告	2359	2	1557	6	(381,1978)
3	Spambase	4601	2	57	3	(2788,1813)
4	图像分割	2310	7	19	2	(330,330,...,330)
5	心电图	2126	3	21	3	(1655,295,176)

多特征数据集是从荷兰实用程序图的集合中提取的手写数字数据集，其中包含 2000 个属于 10 类(0-9)的数字对象。每类有 200 个对象。每个对象均由 649 个特征表示，这些特征分为以下六个视图：a)Mfeat-fou 视图包含 76 个字符形状的傅里叶系数；b)Mfeat-fac 视图包含 216 个配置文件相关性；c)Mfeat-kar 视图包含 64 个 Karhunen-Love 系数；d)Mfeat-pix 视图包含 240 个像素窗口；e)Mfeat-zer 视图包含 47 个 Zernike 时刻；f)Mfeat-mor 视图包含 6 个形态特征。

互联网广告数据集包含来自各种网页的 3279 张图像，这些图像被分类为广告或非广告。有 20 张图片的值缺失。本文的实验在 3259 个实例上进行，删除了缺失值的实例。在六个视图中描述了实例。视图 1 包含 3 种图像几何形状(宽度，高度，长宽比)；视图 2 在包含图片的页面网址(基本网址)中包含 457 个词组；视图 3 包含 495 个图像 URL 的短语(图像 URL)；视图 4 在图像所指向的页面的 URL 中包含 472 个短语(目标 URL)；视图 5 包含 111 个锚文本；视图 6 包含 19 个图像的文本 alt(替代)html 标签(alt 文本)。

Spambase 数据集是一个数据集，其垃圾邮件的收集来自邮局主管和具有现场垃圾邮件的个人，非垃圾邮件的收集来自现场工作和个人电子邮件，其中包含 4601 个属于 2 类(垃圾邮件、非垃圾邮件)的对象。每个对象都由 57 个要素表示，这些要素分为三个视图，分别是单词频率视图，字符频率视图和大写游程视图。a) 单词频率视图包含 word 类型的 48 个连续实数属性；b)字符频率视图包含 char 类型的 6 个连续实数属性；c)大写游程视图包含测量连续大写字母序列长度的 3 个连续实数属性。

在该数据集中，从 7 张室外图像的数据库中随机抽取了 2310 个实例。手动分割图像以为每个像素创建一个分类。每个实例都是一个 3×3 的区域。数据集包含 19 个特征，可分为 2 个视图：形状视图包含 9 个有关形状信息的特征，而 RGB 视图包含 10 个有关颜色信息的特征。

自动处理 2126 例胎儿心电图(Cardiotocograms,CTG)并测量相应的诊断特征。CTG 还由三位专家产科医生进行分类，并为他们每个人分配了共识分类标签。分类既涉及形态学模式(A，B，C ...)，也涉及胎儿状态(N，S，P)。因此，该数据集可用于 10 类或 3 类实验。在此实验中，将其用作 3 类数据集。在数据集中，可以将 21 个要素划分为 3 个视图：每秒指标，可变性视图和直方图视图。

3.3　评估指标

由于已为本文的实验选择了五个数据集的真实分区，因此可以通过将所得聚类与外部结构按照外部标准进行比较来评估聚类算法的性能。一些常用的标准包括兰德指数(Rand Index,RI)，杰卡德系数(Jaccard Coefficient,JC)和 Folk(Fowlkes Russel)。令 $C=\{C_1,C_2,C_M\}$ 为数据集中的 M 个簇，$C'=\{C'_1,C'_2,\ldots,C'_N\}$ 由聚类算法生成的 N 个聚类的集合。给定数据集中的一对点 (X_i,X_j)，称为

SS 是一对数据点的数量，其中 $X_i,X_j\in C_m$，$X_i,X_j\in C'_n,i\neq j$。

DD 是一对数据点的数量，其中 $X_i\in C_{m1},X_j\in C_{m2}$，$X_i\in C'_{n1}$，$X_j\in C'_{n2}$，$i\neq j,m1\neq m2,n1\neq n2$。

SD 是一对数据点的数量，其中 $X_i,X_j\in C_m,X_i\in C'_{n1},X_j\in C'_{n2},i\neq j,n1\neq n2$。

DS 是一对数据点的数量，其中 $X_i\in C_{m1},X_j\in C_{m2}$，$X_i$，$X_j\in C'_n$，$i\neq j,m1\neq m2$。

本文使用的三个外部标准可以定义如下：

$RI=(SS+DD)/(SS+SD+DS+DD)$，RI 值越大，说明聚类结果与真实情况越吻合。

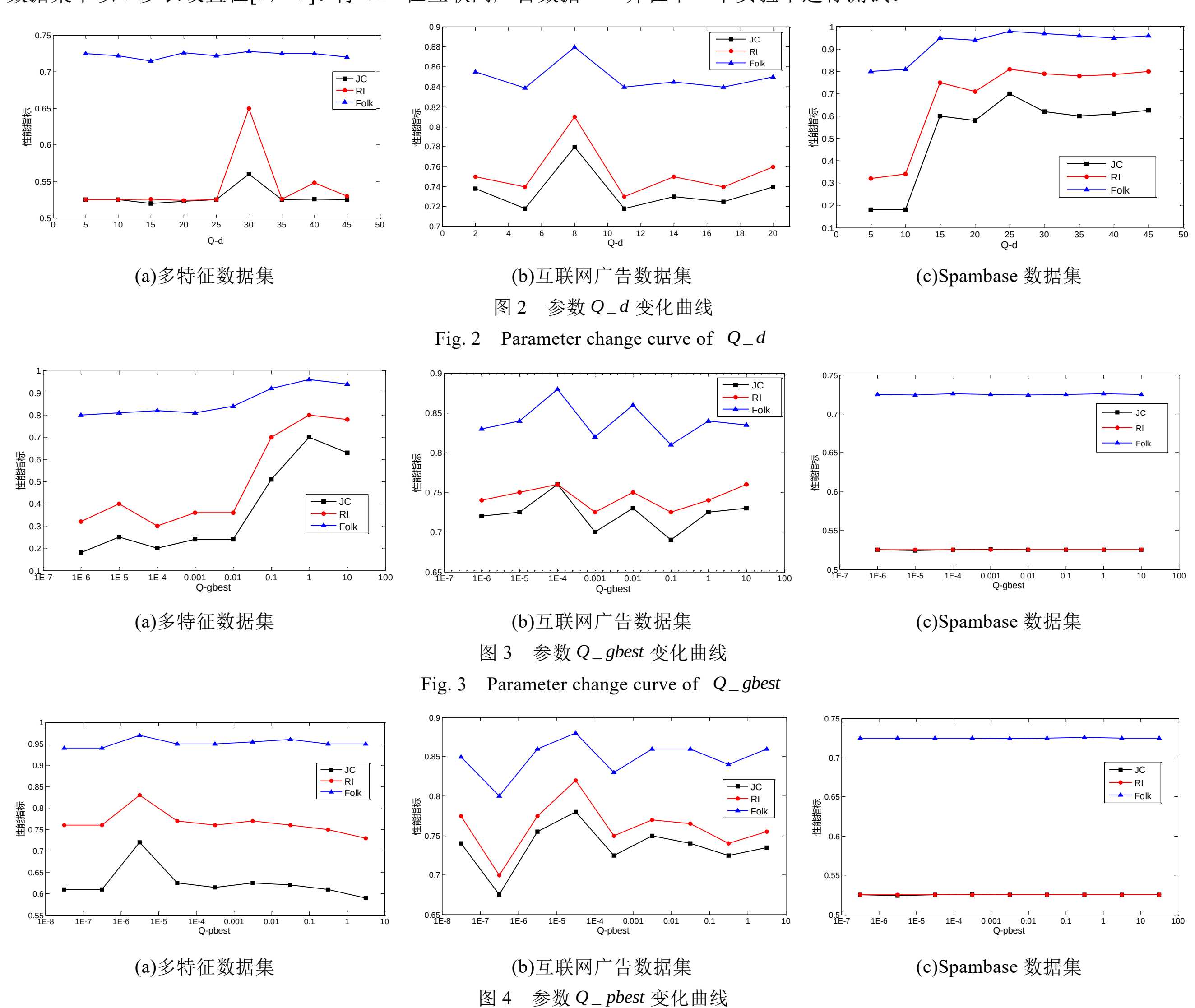
$JC = SS/(SS+SD+DS)$，该指标用于衡量两个数据的相似度，JC 值越大，相似度越大，聚类精度越高。

$Folk = \sqrt{\frac{SS}{SS+SD}} \times \sqrt{\frac{SS}{SS+DS}}$，该指标用于评价聚类质量，Folk 值越大，说明聚类质量越高。

3.4　参数分析

为了分析 3 个参数(Q_d，Q_gbest 和 Q_pBest)对高维多视图数据的聚类性能的影响，在 single node 上对 3 个数据集(Mfeat 数据集，互联网广告数据集以及 Spambase 数据集)中的 IWKM 进行了测试。为了减少统计错误，所有数据集均独立进行了 10 次模拟。

根据过早收敛的阈值，本文将 Q_d 在 Mfeat 和 Spambase 数据集中以 5 步长设置在[5，45]。将 Q_d 在互联网广告数据集中以 3 步长设置在[2，20]。关于它们的平均评价指标的统计结果如图 2 所示。从图 2 可以看出，当 Q_d 分别选择为 25、8 和 30 时，IWKM 具有在 single node 上的 3 个数据集中进行聚类的最佳性能。参数 Q_gbest 和 Q_pBest 是维度惯性的阈值，用于测量每个维度中位置的可感知变化是否发生。三个数据集上的参数 Q_gbest 和 Q_pBest 也与 Q_d 类似地进行分析。关于它们的平均评价指标的统计结果分别示于图 3 和图 4。当参数 Q_gbest 设置为 5,5.0E-5 和 5.0E-4，且参数 Q_pBest 设置为 3.0E-6，3.0E-5 和 0.03，IWKM 在 single node 上的 3 个数据集中的聚类性能是最好的。由于在 Spambase 中 JC 和 RI 的值几乎相等，因此 JC 和 RI 的曲线重叠。因此，根据参数分析的结果，将选择最佳参数值 Q_d，Q_gbest 和 Q_pBest 并在下一个实验中进行测试。



(a)多特征数据集　(b)互联网广告数据集　(c)Spambase 数据集

图 2　参数 Q_d 变化曲线

Fig. 2　Parameter change curve of Q_d

(a)多特征数据集　(b)互联网广告数据集　(c)Spambase 数据集

图 3　参数 Q_gbest 变化曲线

Fig. 3　Parameter change curve of Q_gbest

(a)多特征数据集　(b)互联网广告数据集　(c)Spambase 数据集

图 4　参数 Q_pbest 变化曲线

Fig. 4　Parameter change curve of Q_pbest

3.5　PSO 和 CPSO 的比较

为了验证 CPSO 在 IWKM 中聚类中心、视图权重和特征权重的优化，本文在 single node 上的三个高维多视图数据集中测试了 CPSO 和 PSO。通过 CPSO 和 PSO 将数据集运行 10 次，图 7 记录并比较了各种算法的平均结果。在 CPSO 中，提出了一种精确的扰动，包括合适的扰动粒子、精确的扰动时间和扰动维数，以提高优化性能。从图 7 中可以看出，CPSO 可以在 single node 上的所有三个高维多视图数据集中实现更好的解决方案精度，并尽早获得最佳解决方案。显然，CPSO 在 IWKM 集群方面比 PSO 具有更好的性能。因此，本文可以得出结论，作为一种重要的优化方法，CPSO 可以帮助 IWKM 在高维多视图数据中获得更好的初始聚类中心、视图权重和特征权重。

3.6　IWKM 视图权重比较

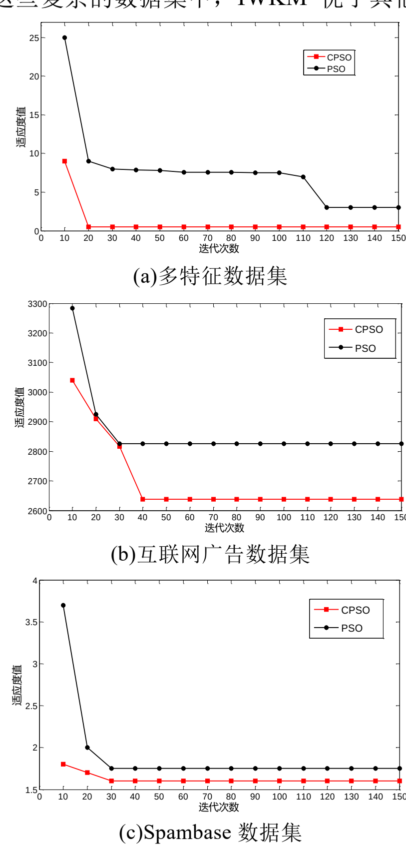
为了进一步评估获得视图权重的性能，在五个不同的高维多视图数据集中测试了 TWKM 和 IWKM。两个算法在数据集中运行了 10 次，并记录了 IWKM 和 TWKM 的平均结果并在表 4 中进行了比较。显然，IWKM 和 TWKM 可以为 5 个高维多视图数据集获得有效权重。特别是，在互联网广告和图像分割这两个数据集中，TWKM 和 IWKM 在获得视图权重方面具有相似的性能。但是，在 apache spark 和 single node 上，在其他 3 个数据集(Mfeat，Spambase 和心电图)中，IWKM 可以获得比 TWKM 更好、更合理的视图权重。TWKM 计算出的视图权重常常集中在一个视图上，这与现实应用不

符。IWKM 计算的权重比 TWKM 计算的权重更合理，并且特征的权重处于相同情况。因此，本文可以得出结论，在视图权重方面，IWKM 比 TWKM 具有更好的性能。

利用 CPSO 进行优化得到六个聚类算法的最优参数值，如表 2 所示。为了进一步验证所提出算法在大数据应用中对高维多视图数据进行聚类的综合性能，在 apache spark 和 single node 两种不同的计算平台上，通过 RI，JC 和 Folk 的评估指标，在五个高维多视图数据集中将 IWKM 与其他五种算法进行了比较。

在实验中，视图数与特征数的乘积记录为 $p_{f\times v}$，用于描述高维多视图数据的复杂性。特征数越大，高维多视图数据越复杂。在表 1 中，根据 $pro_{f\times v}$ 的值，Mfeat 的数据集(特征数：649，视图数：6，$p_{f\times v}=649\times 6=3894$)、互联网广告数据集(特征数：1557，视图数：6，$p_{f\times v}=1557\times 6=9342$)比 Spambase 数据集(特征数：57，视图数：3，$p_{f\times v}=57\times 3=171$)、图像分割数据集(特征数：19，视图数：2，$p_{f\times v}=19\times 2=38$)、心电图数据集(特征数：21，视图数：3，$p_{f\times v}=21\times 3=63$)更复杂。

表 3 总结了 IWKM 与其他 5 种算法在 apache spark 和 single node 上的综合比较。比较它们的平均结果(10 倍)和标准偏差以减少统计误差。从这些结果中，可以看到 IWKM 在 Mfeat 数据集和互联网广告数据集中明显优于的其他五种算法。在 Spambase 数据集中，IWKM 的性能优于 TWKM 和 DC，但 AP 在 Mfeat 数据集中的效果最差。在 Mfeat 数据集中，DC 和 IWKM 均比 LAC，AP，Ncut 和 TWKM 更好。在互联网广告数据集中，AP，TWKM 和 IWKM 的性能优于 LAC，Ncut 和 DC。LAC 明显优于 Spambase 数据集中的其他 5 种算法(包括 IWKM)，但 Spambase 数据集的复杂度低于 Mfeat 数据集和互联网广告数据集。因此，可以得出结论，在这些复杂的数据集中，IWKM 在针对具有更多视图和更高维度数据集来说，例如多特征和互联网广告数据集，胜过其他五种算法。在心电图数据集中，IWKM 优于其他的 5 种算法。但是，在图像分割中，Ncut 和 TWKM 的性能要优于 IWKM。由于心电图数据集比图像分割数据集更为复杂，因此高维多视图数据集越复杂，IWKM 的性能越好。总之，IWKM 可以更加有效地处理大数据应用中的高维多视图数据集的聚类。同时，在这些复杂的数据集中，IWKM 优于其他五种算法。



(a)多特征数据集

(b)互联网广告数据集

(c)Spambase 数据集

图 5　PSO 与 CPSO 的比较

Fig. 5　Comparison between PSO and CPSO

表 2　实验中六种聚类算法的参数值

Tab. 2　Parameter values of six clustering algorithms in the experiment

算法	多特征	互联网广告	Spambase	图像分割	心电图
LAC(h)	2	2	14	2	5
AP(λ， p)	(0.9,2.7)	(0.9,60.0)	(0.9,12.0)	(0.9,-4.7)	(0.9,-24)
Ncut(e)	1.0E-8	1.0E-8	1.0E-8	1.0E-8	1.0E-8
DensityC(p)	1.6	1.4	1.9	1.7	1.5
TWKM(λ,η)	(30,7)	(80,25)	(53,18)	(70,40)	(40,18)
IWKM(Q_d, Q_gBest, Q_pBest)	(25.0,0.5,3.0E-6)	(8.0,5.0E-5,3.0E-5)	(30.0,5.0E-4,0.03)	(20.0,5.0,3.0E-5)	(20,5.0,3.0)

表 3　五种算法的比较

Tab. 3　Comparison of five algorithms

数据集		LAC	AP	Ncut	DensityC	TWKM	IWKM
多特征	RI	0.9344 ±0.0000	0.8931 ±0.0000	0.9317 ±0.0065	0.9578 ±0.0000	0.9456 ±0.0000	0.9586 ±0.0118
	JC	0.5365 ±0.0000	0.3510 ±0.0000	0.4959 ±0.0342	0.6720 ±0.0000	0.5937 ±0.0000	0.6820 ±0.0719
	Folk	0.6988 ±0.0000	0.5226 ±0.0000	0.6625 ±0.0301	0.8060 ±0.0000	0.7467 ±0.0000	0.8116 ±0.0466
互联网广告	RI	0.7154 ±0.0000	0.8124 ±0.0000	0.6803 ±0.0016	0.6996 ±0.0000	0.8131 ±0.0000	0.8179 ±0.0132
	JC	0.7055 ±0.0000	0.7785 ±0.0000	0.6151 ±0.0026	0.6974 ±0.0000	0.7792 ±0.0000	0.7858 ±0.0088
	Folk	0.8322 ±0.0000	0.8759 ±0.0000	0.7646 ±0.0017	0.8293 ±0.0000	0.8764 ±0.0000	0.8809 ±0.0043
Spambase	RI	0.7112 ±0.0000	0.5527 ±0.0000	0.5616 ±0.0000	0.5209 ±0.0000	0.5208 ±0.0000	0.5225 ±0.0003
	JC	0.5893 ±0.0000	0.4797 ±0.0000	0.4611 ±0.0000	0.5196 ±0.0000	0.5194 ±0.0000	0.5222 ±0.0002
	Folk	0.7397 ±0.0000	0.6590 ±0.0000	0.6358 ±0.0000	0.7194 ±0.0000	0.7192 ±0.0000	0.7225 ±0.0003
图像分割	JC	0.3252 ±0.0000	0.2254 ±0.0000	0.3038 ±0.0000	0.2573 ±0.0000	0.2996 ±0.0000	0.2297 ±0.0065
	RI	0.5319 ±0.0000	0.8110 ±0.0000	0.8974 ±0.0000	0.5388 ±0.0000	0.8252 ±0.0000	0.8047 ±0.0103
	Folk	0.5055 ±0.0000	0.3682 ±0.0000	0.4706 ±0.0000	0.4115 ±0.0000	0.4645 ±0.0000	0.3750 ±0.0036
心电图	JC	0.3854 ±0.0000	0.3067 ±0.0000	0.1886 ±0.0000	0.3535 ±0.0000	0.3897 ±0.0000	0.3984 ±0.0000
	RI	0.5408 ±0.0000	0.5034 ±0.0000	0.4346 ±0.0000	0.4617 ±0.0000	0.5086 ±0.0000	0.5576 ±0.0210
	Folk	0.5705 ±0.0000	0.4885 ±0.0000	0.3721 ±0.0000	0.5262 ±0.0000	0.5656 ±0.0000	0.5854 ±0.0054

表 4　TWKM 和 IWKM 计算的视图权重

Tab. 4　View weights calculated by TWKM and IWKM

	TWKM 特征权值	IWKM 特征权值
多特征	1.66665E −6	0.23424
	1.66665E −6	0.23358
	1.66665E −6	0.25141
	1.66665E −6	0.01263
	1.66665E −6	0.09903
	0.99999	0.16911
互联网广告	1.66665E −6	0.11030
	0.20205	0.16166
	0.21539	0.12580
	0.19255	0.29347
	0.16216	0.30720
	0.22784	0.00157
Spambase	0.99999	0.58757
	3.33331E −6	0.06495
	3.33331E −6	0.34748
图像分割	0.4684598	0.44744640
	0.5315402	0.55255359
心电图	9.999933E−01	0.1592640
	3.333311E−06	0.4687741
	3.333311E−02	0.3719617

4　结束语

针对传统聚类算法无法处理大数据中多视图高维数据问题，提出了一种基于混沌粒子群优化算法的智能加权 K 均值聚类算法。通过实验证明了 CPSO 可以帮助 IWKM 在高维多视图数据中获得更好的初始聚类中心、视图权重和特征权重，为聚类精度的提升提供良好的初始值要求。另外提出方法能够有效实现多视图高维数据的聚类，且针对视图越多、维数越高、数据越复杂的数据集越能够体现该算法的优越性。但是本文方法由于数据来源问题，只应用了五类数据，对方法的验证效果还需要更多类别的数据进行验证，需要进一步研究。

参考文献：

[1] 臧艳辉，赵雪章，席运江. Spark 框架下利用分布式 NBC 的大数据文本分类方法 [J]. 计算机应用研究, 2019, 36 (12): 3705-3708+3712. (Zang Yanhui, Zhao Xuezhang, Xi Yunjiang. Large data text classification using distributed NBC in spark framework [J]. Computer application research, 2019, 36 (12): 3705-3708+3712.)

[2] 邹劲松，李芳. 大数据下的分布式精确模糊 KNN 分类算法 [J]. 计算机应用研究, 2019, 36 (12): 3701-3704. (Zou Jinsong, Li Fang. Distributed accurate fuzzy KNN classification algorithm under big data [J]. Computer application research, 2019, 36 (12): 3701-3704.)

[3] 张贝娜，冯震华，张丰，等. 基于时空多视图 BP 神经网络的城市空气质量数据补全方法研究 [J]. 浙江大学学报（理学版）, 2019, 46 (06): 737-744. (Zhang Beina, Feng Zhenhua, Zhang Feng, *et al.* Study on the method of urban air quality data completion based on spatiotemporal multi view BP neural network [J]. Journal of Zhejiang University (SCIENCE EDITION) , 2019, 46 (06): 737-744.)

[4] Shi Hong, Li Yan, Han Yang, et al, Cluster structure preserving unsupervised feature selection for multi-view tasks, Neurocomputing 175 (2016) 686–697.

[5] 张天真. 基于非负矩阵分解的多视图聚类方法研究 [D]. 西安电子科技大学, 2018. (Zhang Nai. Multi view clustering based on nonnegative matrix decomposition [D]. Xi'an University of Electronic Science and technology, 2018.)

[6] Li Hang, He Hong, Wen ying, Dynamic particle swarm optimization and k-means clustering algorithm for image segmentation, Opt. -Int. J. Light Electron Opt. 2018, 126 (24): 4817-4822.

[7] 洪敏，贾彩燕，李亚芳，等. 样本加权的多视图聚类算法 [J]. 计算机研究与发展, 2019, 56 (08): 1677-1685. (Hong Min, Jia Caiyan, Li Yafang, *et al.* Sample weighted multi view clustering algorithm [J]. Computer research and development, 2019, 56 (08): 1677-1685.)

[8] 梁丹，于海燕，范九伦，等. 核空间局部自适应模糊 C-均值聚类图像分割算法 [J]. 微电子学与计算机, 2019, 36 (02): 21-25. (Liang Dan, Yu Haiyan, fan Jiulun, *et al.* Kernel space local adaptive fuzzy c-means clustering image segmentation algorithm [J]. Microelectronics and computer, 2019, 36 (02): 21-25.)

[9] Chen Xu, Xu Xuan, Huang Jiazhi, *et al.* Tw-k-means: automated two-level variable weighting clustering algorithm for multiview data, IEEE Trans. Knowl. Data Eng. 2016, 28 (4): 932-944.

[10] Kumar D, Bezdek J, Palaniswami M, et al, A hybrid approach to clustering in big data, IEEE Trans. Cybern. 2016, 46 (10): 2372–2385.

[11] 王杰，陈彬，袁鹏，等. 数据驱动的最优互惠避碰模型偏好速度研究 [J]. 系统仿真学报, 2019, 31 (12): 2731-2739. (Wang Jie;Chen Bin, Yuan Peng, *et al.* Study on data-driven orca preference velocity [J]. Journal of System Simulation, 2019, 31 (12): 2731-2739.)

[12] 倪龙强，张丽华，姚新涛，等. 一种基于粗糙集证据理论深度融合的局部冲突快速合成方法 [J]. 兵工学报, 2019, 40 (12): 2560-2569. (Ni Longqiang, Zhang Lihua, Yao Xintao, *et al.* A fast synthesis method of local conflicts based on deep fusion of rough set evidence theory [J]. Journal of military engineering, 2019, 40 (12): 2560-2569.)

[13] 范虹，侯存存，朱艳春，等. 烟花算法优化的软子空间 MR 图像聚类算法 [J]. 软件学报, 2017, 28 (11): 3080-3093. (Fan Hong, Hou Cuncun, Zhu Yanchun, *et al.* Soft subspace MR image clustering algorithm optimized by fireworks algorithm [J]. Journal of software, 2017, 28 (11): 3080-3093.)

[14] 王彩云，黄盼盼，李晓飞，等. 基于 AEPSO-SVM 算法的雷达 HRRP 目标识别 [J]. 系统工程与电子技术, 2019, 41 (09): 1984-1989. (Wang Caiyun, Huang Panpan, Li Xiaofei, *et al.* Radar HRRP target recognition based on aepso-svm algorithm [J]. System engineering and electronic technology, 2019, 41 (09): 1984-1989.)

[15] 刘久富，丁晓彬，郑锐，等. 混沌量子粒子群的权重类条件贝叶斯网络分类器参数学习 [J]. 系统工程与电子技术, 2019, 41 (10): 2304-2309. (Liu Jiufu, Ding Xiaobin, Zheng Rui, *et al.* Parameter learning of Bayesian network classifier with weight class condition of Chaos Quantum Particle Swarm [J]. System engineering and electronic technology, 2019, 41 (10): 2304-2309.)

[16] Tang Zi, Liu Min, Ammar A, *et al.* An optimized mapreduce workflow scheduling algorithm for heterogeneous computing, J. Supercomput. 2016, 72 (6): 2059–2079.

[17] 陈云芳，夏涛，张伟，等. 基于亲和传播的动态社会网络影响力扩散模型 [J]. 通信学报, 2016, 37 (10): 40-47. (Chen Yunfang, Xia Tao, Zhang Wei, *et al.* Dynamic social network influence diffusion model based on affinity Communication [J]. Journal of communications, 2016, 37 (10): 40-47)

[18] 陈宸，叶波，邓为权，等. 基于 SLIC 超像素算法和密度聚类的 TA2 钛板表面缺陷定量化评估研究 [J]. 电子测量与仪器学报, 2019, 33 (11): 128-135. (Chen Chen, Ye Bo, Deng Weiquan, *et al.* Quantitative evaluation of surface defects of TA2 titanium plate based on SLIC super-pixel algorithm and density clustering [J]. Journal of electronic measurement and instrumentation, 2019, 33 (11): 128-135.)